MASTER COPY — RODUCTION PURPOSES

AD-A238 097

REPORT DOCU

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 25, 1991	Final Report 1 Jul 87 - 31 Jan 91

4. TITLE AND SUBTITLE
Hyperdimensional Data Analysis and Structural Inference

5. FUNDING NUMBERS
DAAL03-87-K-0087

6. AUTHOR(S)
Edward J. Wegman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
George Mason University
4400 University Drive
Fairfax, VA 22030

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 24105.19-MA

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project focused on exploratory and structural inference for high dimensional data. Exploratory techniques focused mainly on the parallel coordinate technique which was exploited on graphical computing platforms. A MS-DOS package was developed using this and other dynamical graphics techniques. Several refinements of the parallel coordinate technique were made including scintillation, parallel coordinate density plots and grand tours in d-dimensions. The structural inference work is nonparametric in character and focuses on high dimensional density estimation and ridge estimation. These ideas are currently still under development and promising techniques involved random tessellations, the trajectory method for finding ridges and the gradient density plot.

14. SUBJECT TERMS
scintillation, parallel coordinate density plots, grand tours, random tessellations, trajectory method, gradient density plots

15. NUMBER OF PAGES
10

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Hyperdimensional Data Analysis and Structural Inference

Edward J. Wegman

Final Technical Report

May 25, 1991

U. S. Army Research Office

DAAL03-87-K-0087

George Mason University

**Approved for Public Release;
Distribution Unlimited**

Abstract

This project focused on exploratory and structural inference for high dimensional data. Exploratory techniques focused mainly on the parallel coordinate technique which was exploited on graphical computing platforms. A MS-DOS package was developed using this and other dynamical graphics techniques. Several refinements of the parallel coordinate technique were made including scintillation, parallel coordinate density plots and grand tours in d-dimensions. The structural inference work is nonparametric in character and focuses on high dimensional density estimation and ridge estimation. These ideas are currently still under development and promising techniques involved random tessellations, the trajectory method for finding ridges and the gradient density plot.

Table of Contents

1. Introduction.

Electronic instrumentation implies an ability to acquire a large amount of high dimensional data very rapidly. While such capabilities have existed for some time, the emergence of cheap RAM in the 1980s has given us the ability to store and access that data in an active computer memory. It is our thesis that this new ability represents a challenge for statisticians to develop methodology that is substantially different in kind from previous methodology. Historically, the majority of existing methodology is focused on the univariate, i.i.d. random variable model. Even in the circumstance that a multivariate model is allowed, it is usually assumed to be multivariate normal. A premise of our research has also been that while arbitrary sample size is frequently assumed, the truth of the matter is that common techniques implicitly assume small to moderate sample sizes. For example a regression problem with 5 design variables and 1000 observations would represent no problem for traditional techniques. By contrast a regression problem with 40,000 design variables and 8 million observations would. The reason is clear. In the former case the emphasis is on statistical efficiency which implicitly most current statistical methodology optimizes. By contrast the latter case emphasis must clearly be on computational efficiency. The emphasis placed on parsimony in many contemporary books and papers is a further reflection of the mentality focused on small to moderate sample sizes. Finally, we note that the very fact of largeness in sample size implies that it is unlikely we would see i.i.d. homogeneity.

The premise of this research project was to focus on the new perspective required for the analysis of large, high dimensional data sets. We intended to initiate an exploration of this perspective. Primary focus was on data representation and data exploration and secondly on structural inference. The tasks related to these two foci were are outlined below.

1. Develop techniques for visualizing traditional statistical notions such as clustering, correlation, symmetry in parallel coordinate displays.

2. Relate statistical measures to simple statistical features of the parallel coordinate display.

3. Explore hyperdimensional geometry in parallel coordinates.

4. Implement parallel coordinate diagrams on mini/micro computers.

5. Construct the grand tour in parallel coordinates.

6. Explore the combinatorial issues of parallel coordinates.

7. Develop a high speed algorithm for multidimensional density estimation based on the Dirichlet tessellation

8. Develop the theoretical properties of this multidimensional density estimate.

9. Develop a formal definition of and estimation procedures for d-ridges.

10. Develop a parallel computing algorithm for multidimensional density estimation and d-ridge estimation. Develop the corresponding graphical tools. Develop the associated non-Gaussian statistical inference theory.

2. Summary of Results

a. Parallel Coordinate Displays. The parallel coordinate multidimensional data representation has been studied thoroughly. The parallel coordinate display is an excellent tool for statistical data representation as it is shown to be characterized by projective transformations of E^n into E^2. This feature gives an elegant matrix formulation of parallel coordinate diagrams and also assures that the duality properties commonly found in projective geometry hold for parallel coordinate displays. As an example, quadratic forms map into quadratic forms, i.e. conic sections map into conic sections. One particularly useful duality is that ellipses map into hyperbolas. This has a very useful interpretation for data which has a joint density with ellipsoidal cross sections. Hyperdimensional ellipsoids are easily recognized in parallel coordinates. We have found particularly useful ideas in working with parallel coordinate displays to be implementations of painting (brushing), scintillation, density plots and grand tours.

As with all data plots having a large number of observations, there is heavy overplotting in parallel coordinate displays. We have improved the situation substantially by using different colors for plotting each observation. To a large extent this allows us to track a single observation through all dimensions of a parallel coordinate display. By using various colored paints including rainbow paint and invisible paint, we have been able to manipulate the displays very effectively. Solid colored paint allows us to paint clusters and separate blocks of data in a natural way. Invisible paint allows us to eliminate outlier and/or subclusters to examine the regular structure more carefully. Rainbow paint allows us to return to multicolored diagrams easily.

When a diagram is rainbow painted and there is still a considerable amount of overplotting, we have found scintillation to be a useful tool. The idea is that if each

observation is a different color and we simply swap colors sequentially at all points where there is overplotting, we obtain a dynamical flashing display in which the speed of flashing conveys a sense of how many points are overplotted. The dynamic color display also allows us to track observations where there would be ambiguity due to overplotting.

An alternate technique when overplotting is even heavier is to construct the parallel coordinate (line) density display. The idea is to replace the parallel coordinate diagram with a density version of it. This has enormous benefit when global structure is obscured by overplotting even when the sample sizes are fairly small. We have demonstrated ability to detect a hole on the inside of a 4-dimensional hypersphere when we have only sampled data. This is an excellent achievement because 1. the spherical symmetry of a 4-sphere eliminates any preferential projection direction and 2. any three or lower dimensional projection of a 4-sphere will not show any holes.

The final idea in connection with parallel coordinates that we have explored is the grand tour. The grand tour tries to capture the idea of looking at the data from all possible directions. It may be thought of as a general affine rotation of a d-dimensional coordinate system in d-space and then plotting data points expressed in the rotated coordinate system in a screen display parallel coordinate system. This has proven extraordinarily effective and has allowed us to demonstrate finding linear substructures of 6 dimensions in 9-dimensional data and also to find 7-dimensional clusters in 9-dimensional data.

b. Other Results. A wide variety of other items were addressed. Some of the tasks have been addressed but are not yet in technical report form. I will mention some of these results briefly. We have now formal definition of d-ridges and have several methods for estimation of d-ridges. One of the most interesting is a direct trajectory method based on multidimensional density hypersurfaces. A two-dimensional illustration is given in Figure 1. The ridge is apparent in this display. The location of the ridge seems fairly robust with respect to the smoothing parameters. In order to locate the ridge more precisely, we have constructed a gradient based technique. The estimated gradient is shown in a 3-d perspective plot in Figure 2. Figures 1 and 2 are based on the same data. It should be apparent that these are highly nonlinear data structures and we have quite successful estimates of them. The ideas generalize in a straightforward way to general d-dimensional space. This work is being carried out with my Ph.D. student Qiang Luo. Our work on density estimates based random

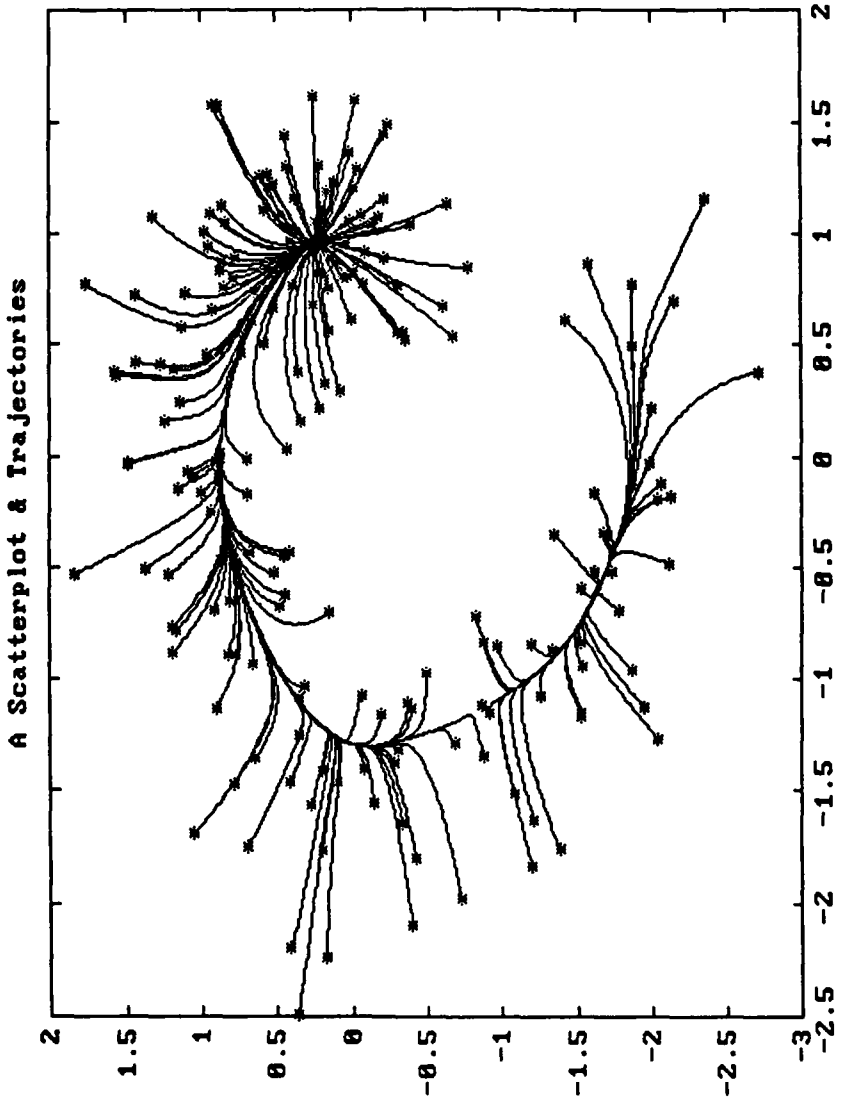

Figure 1. Nonlinear curve (C-curve) estimated using trajectory method based a smoothed density plot.

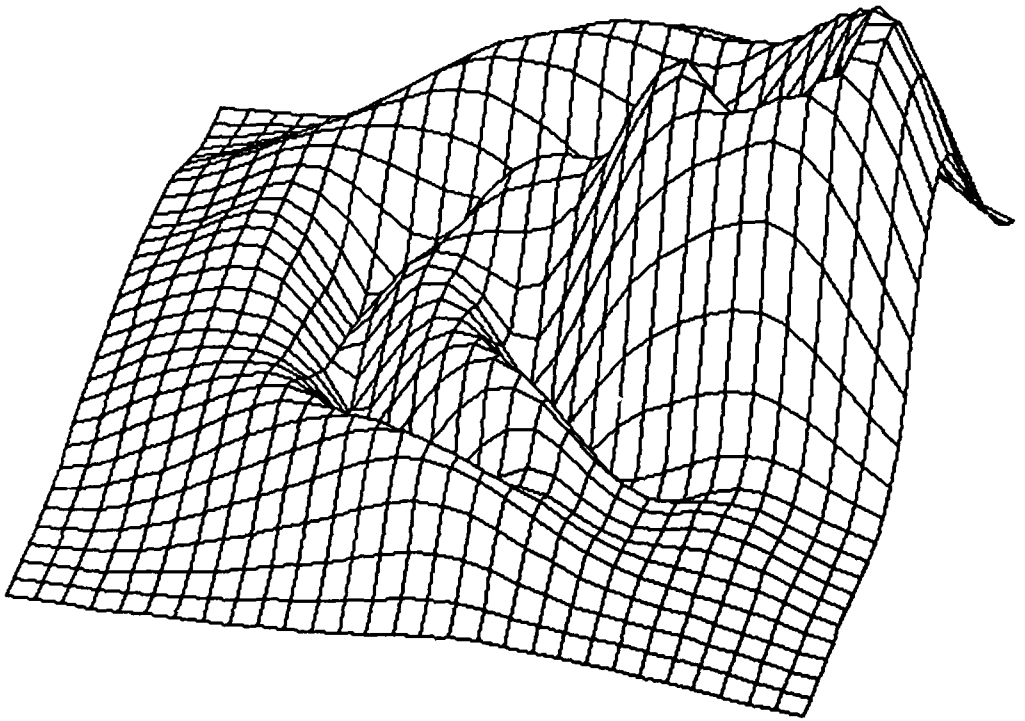

Figure 2. 3-dimensional Perspective Plot of the Density Derivative of the Nonlinear Curve (C-curve).

tessellations is also nearing completion and will be reported on soon. That work is being carried out with my Ph.D. student Leonard Hearne. All of the tasks listed above and in our 1987 proposal have been carried out. A complete list of reports follows.

3. List of Publications and Technical Reports

John J. Miller and Edward J. Wegman, "Vector function estimation using splines," ***J. Statist. Plan. Infer.,*** 17, 173-180, 1987

Edward J. Wegman, "Reproducing kernel Hilbert spaces," ***Encyclopedia of Statistical Sciences,*** (N. Johnson, S. Kotz and C. Read, eds.), 8, 81-84, John Wiley and Sons: New York, 1988

Edward J. Wegman, "Sobolev spaces," ***Encyclopedia of Statistical Sciences,*** (N. Johnson, S. Kotz and C. Read, eds.), 8, 535-537, John Wiley and Sons: New York, 1988

Edward J. Wegman and Annie Hayes, "Statistical software," ***Encyclopedia of Statistical Sciences,*** (N. Johnson, S. Kotz and C. Read, eds.), 8, 667-674, John Wiley and Sons: New York, 1988

Edward J. Wegman, "A view of computational statistics and its curriculum," ***Am. Statist. Assoc. Proc. Sect. Statist. Educat.*** 1-6, 1988

Edward J. Wegman and Herbert Solomon, "Introduction to assessing uncertainty," ***J. Statist. Planning Infer.,*** 20, 241-244, 1988

Edward J. Wegman, "On randomness, determinism and computability," ***J. Statist. Planning Infer.,*** 20, 279-294, 1988

Edward J. Wegman, "Computational statistics: a new agenda for statistical theory and practice," ***J. Washington Acad. Science,*** 78, 310-322, 1988.

Masood Bolorforoush and Edward J. Wegman, "On some graphical representations of multivariate data," ***Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface***, 121-126, 1988.

Edward J. Wegman, Invited discussion of "How to get your first research grant," by B. E. Trumbo, ***Statistical Science,*** 4(2), 146-148, 1989.

Edward J. Wegman, "Parallel coordinate densities," ***Proceedings of the 34th Conference on the Design of Experiments in Army Research Development and Testing***, 247-264, 1989.

Edward J. Wegman, "Parallel computing and statistics," ***Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions,*** 231-244, 1989

Edward J. Wegman, "Stochastic load balancing in parallel computers," ***Proceedings of the Fourth Conference on Hypercubes, Concurrent Computers, and Applications,*** 627-630, 1990

R. Duane King and Edward J. Wegman, "A parallel implementation of data set mapping," ***Proceedings of the Fourth Conference on Hypercubes, Concurrent Computers, and Applications,*** 1197-1200, 1990

Mingxian Xu, John J. Miller and Edward J. Wegman, "Parallelizing multiple linear regression for speed and redundancy: an empirical study," ***Computing Science and Statistics: Proceedings of the 21st Symposium on the Interface,*** 138-144, 1990

John J. Miller and Edward J. Wegman, "Construction of line densities for parallel coordinate plots," ***Computing Science and Statistics: Proceedings of the 21st Symposium on the Interface,*** (short version), 191-199, 1990

Edward J. Wegman, "Hyperdimensional data analysis using parallel coordinates," ***J. American Statist. Assoc.,*** 85, 664-675, 1990

Edward J. Wegman, "Statistics," ***McGraw-Hill Yearbook of Science and Technology: 1991***, 414-416, McGraw-Hill: New York, 1990

John J. Miller and Edward J. Wegman, "Construction of line densities for parallel coordinate plots," in ***Computing and Graphics in Statistics***, (full refereed version), (A. Buja and P. Tukey, eds.), Springer-Verlag: New York, 1991

Edward J. Wegman, "A stochastic approach to load balancing in coarse grain parallel computers," in ***Computing and Graphics in Statistics***, (A. Buja and P. Tukey, eds.), Springer-Verlag: New York, 1991

Melanie Tompkins, ***Testing Linear Hypotheses in Unbalanced Data Designs***, Center for Computational Statistics Technical Report 55, George Mason University, December, 1989 (M. S. Thesis)

Qiang Luo, ***Parity Equation Approach to Failure Detection and Isolation in Dynamical Systems***, Center for Computational Statistics Technical Report 60, George Mason University, April, 1990 (M.S. Thesis)

Edward J. Wegman and Muhammad K. Habib, "Stochastic methods for neural systems," to appear ***Journal of Statistical Planning and Inference*** (1991)

Edward J. Wegman, "The grand tour in k-dimensions," to appear ***Computing Science and Statistics: Proceedings of the 22nd Symposium***, (C. Page, ed.), Springer-Verlag: New York, 1991

Leonard B. Hearne and Edward J. Wegman, "Adaptive probability density estimation in lower dimensions using random tessellations," to appear ***Computing Science and Statistics: Proceedings of the 23rd Symposium***, (E. Keramides, ed.), Springer-Verlag: New York, 1991

Edward J. Wegman, Donald T. Gantz and John J. Miller (eds.), ***Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface,*** Alexandria, VA: American Statistical Association for the Interface Foundation of North America, 1988

Mason Hypergraphics, copyright (c) 1988, 1989 by Edward J. Wegman and Masood Bolorforoush, copyright (c) 1990, 1991 by PSF, Ltd., a MS-DOS package for high-dimensional data analysis

4. List of Participating Personnel

Edward J. Wegman, Dunn Professor of Information Technology and Applied Statistics

John J. Miller, Associate Professor of Applied Statistics

Melanie Tompkins, earned M.S. in Statistical Science

Leonard B. Hearne, earned M.S. in Systems Engineering, admitted to Ph.D. Candidacy

Qiang Luo, earned M.S. in Electrical Engineering, admitted to Ph.D. Candidacy

Mingxian Xu, admitted to Ph.D. Candidacy

Don Faxon, admitted to Ph.D. Candidacy

Masood Bolorforoush, earned M.S. in Systems Engineering

R. Duane King, will earn M.S. in Computer Science, summer, 1990

5. Report of Inventions

None